SCALABLE MEMORY

Abstract

A memory structure and method for handling memory requests from a processor and for returning correspondence responses to the processor from various levels of the memory structure. The memory levels of the memory structure are interconnected by a forward and return path with the return path having twice the bandwidth of the forward path. An algorithm is used to determine how many responses are sent from each memory level on the return path to the processor. This algorithm is designed to guarantee a constant bound on the rate of responses sent to the processor. More specifically, if a write request is at the same level to which it is targeted, or if a request at a memory level is targeted to a higher memory level, then two responses are forwarded from a controller at the memory level on the return path to the processor. Otherwise, only one response is forwarded from the memory level on the return path.